MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2016/2017

TDS 3301 – DATA MINING

(All sections / Groups)

11 MARCH 2017
2:30 p.m – 4:30 p.m
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This question paper consists of **5** pages, inclusive of the cover page, with FOUR questions only.

2. Answer **ALL** questions.

3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

(a) Draw a diagram showing the Knowledge Discovery in Database (KDD) process. Identify the role of data mining within this process, and mention clearly the *input* and the *output* of the data mining process.

(3 marks)

(b) A study was conducted on the effectiveness of *self-quantification,* where students in an all-girl's school were provided with a device which measured their daily steps and hours of sleep. The data from the device was then correlated with the students' body mass index, with an additional variate in the form of number of home-cooked meals taken per week. The data was then captured in the form of a table as shown in Table 1. Body mass index (BMI) is calculated using the following formula : *Bodyweight in kilograms divided by height in meters squared*

Table 1 : Variates related to BMI of students in an all-girl secondary school

Name	Age	Gender	Weight/Kg	Height/m	Body Mass Index	Number of home cooked meals per week	Number of steps per day	Average hours of sleep per day
String	Integer	Binary	Real	Real	Real	Decimal	Integer	Real

I. Which are the variables/columns which can be excluded from the study? Give a reason for each variable/column which you have chosen to be excluded.

(3 marks)

II. What additional **transformation** step on the variables is necessary to perform accurate data mining? (2 marks)

(c) What is the purpose of Principal Component Analysis and what is the criteria for choosing a particular principal component? (2 marks)

Continued...

QUESTION 2

(a) Discuss the differences between classification and prediction, in terms of *what* is being predicted. Give one example for classification and prediction. (3 marks)

(b) Refer to Table 2, which shows the class labels for purchasing decisions made by customers on purchasing a customer. Draw a Decision Tree which shows how the class labels can be potentially decided. The decision tree need not be optimal, and can be based on the algorithm of your choice. (3 marks)

Table 2: Purchasers of computers and their labels

RID	Age	Income	Student	Credit_rating	Class (buy_computer)
1	<=30	High	No	Fair	No
2	<=30	High	No	Excellent	No
3	31...40	High	No	Fair	Yes
4	>40	Medium	No	Fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	Low	Yes	Excellent	No
7	31...40	Low	Yes	Excellent	Yes
8	<=30	Medium	No	Fair	No
9	<=30	Low	Yes	Fair	Yes
10	>40	Medium	Yes	Fair	Yes
11	<=30	Medium	Yes	Excellent	Yes
12	31...40	Medium	No	Excellent	Yes
13	31...40	High	Yes	Fair	Yes
14	>40	Medium	No	Excellent	No

(c) Describe the Greedy algorithm used in inducting a Decision Tree. The algorithm should describe the construction of the decision tree, and continue until the selection of test attributes. (3 marks)

(d) State one weakness of the decision tree algorithm as a classification algorithm. (1 mark)

Continued...

QUESTION 3

(a) Consider Table 2 (same table is used in Question 2) as the training set which we are labeling. Calculate the following for a sample X with the following values: age= youth, income = medium, student = yes, credit =fair. Youths have an age of <= 30. (6 marks)

i. P(X| buy = yes) [Hint: Calculate the probability of all those who fit the given parameters and bought a computer]
ii. P(X| buy = no) [Hint: Calculate the probability of all those who fit the given parameters and did not buy a computer]
iii. What is the predicted class label for sample X ?[Hint: Maximise P(X|Class LabelXP(ClassLabel]

(b) What is an important assumption to make when we attempt to perform a classification task using the Naïve Bayes algorithm and state an advantage of using the Naïve Bayes algorithm. (2 marks)

(c) State TWO ways in which we can improve the performance of a Naïve Bayes model in terms of processing time and model complexity. (2 marks)

Continued...

QUESTION 4

(a) We are living in an age where devices around us are becoming increasingly inter-connected on a real-time basis. Discuss using up to 10 sentences, how stream mining is enabling real-time data mining, especially with respect to the building of models using constantly updated data. (3 marks)

(b) Describe what deep learning is, and explain one use case which has shown the promise of using deep learning. What is the role of graphical processing units in realizing deep learning? (4 marks)

(c) What is the future of data mining? Explain your answer in terms of learning curve, accessibility and automation.

(3 marks)

End of pages.